Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 3

1. Libro Seleccionado — *The Scarlet Letter*

En el presente trabajo se continua con el libro *The Scarlet Letter*, el cual se encuentra disponible de manera gratuita en *Project Gutenberg* [2]. En esta ocasión, se analizó el tipo de distribuciones discretas que podrían estar contenidas en el escrito mediante el programa R versión 4.0.2 [3]. El código en R se encuentra en el repositorio de GitHub [1].

2. Distribuciones

Para el análisis de las distribuciones, se consideró reflejar cómo se comporta la longitud de las palabras presentes en el libro y la cantidad de palabras en cada párrafo. Para este último criterio, se determinó usar párrafos que contienen más de 9 palabras, ya que el título más largo contiene 9 palabras.

De acuerdo a la figura 1, se puede observar que la distribución de la longitud de cada palabra se asemeja a una distribución binomial. Mientras que, en la figura 2 correspondiente a la cantidad de palabras por párrafos, es similar a la geométrica.

3. Simulación de distribuciones

Con el fin de simular las distribuciones geométrica, binomial y binomial negativa, se escogieron como criterios algunas características relevantes del análisis estadístico realizado anteriormente, como por

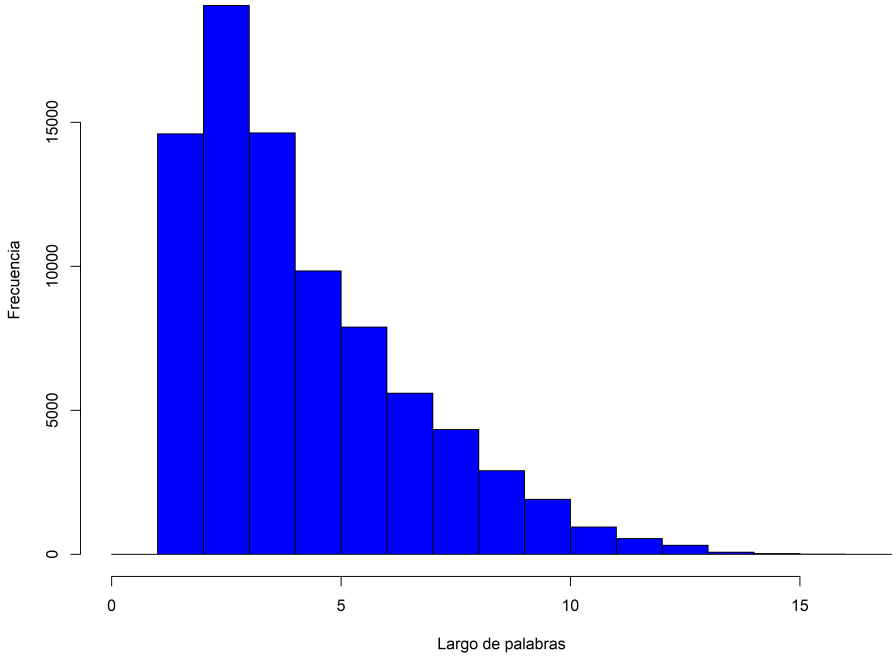

Figura 1: Distribución de la longitud de las palabras usadas en el libro

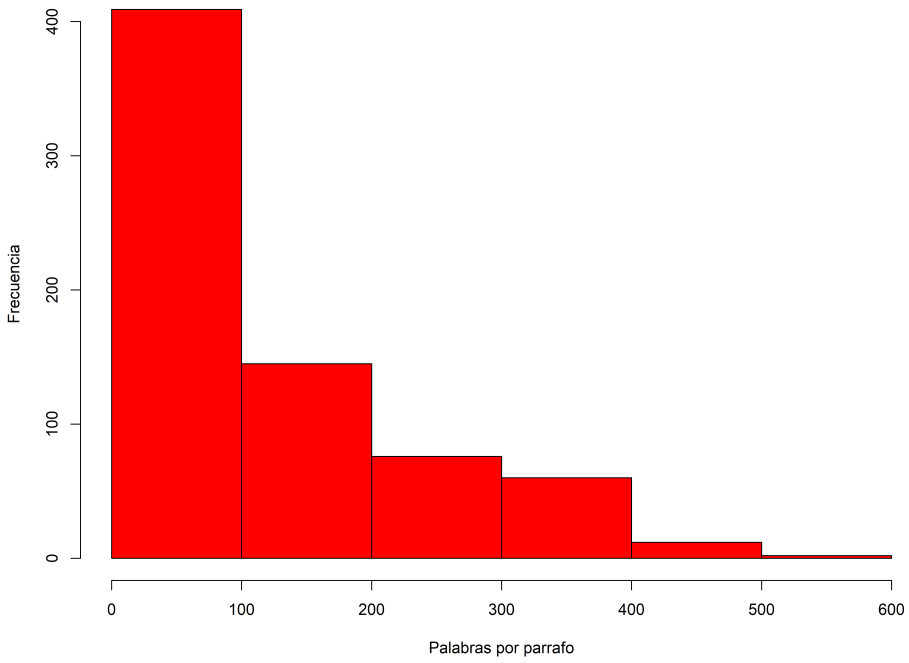

Figura 2: Distribución de la cantidad de palabras en cada párrafo (en párrafos con más de 9 palabras)

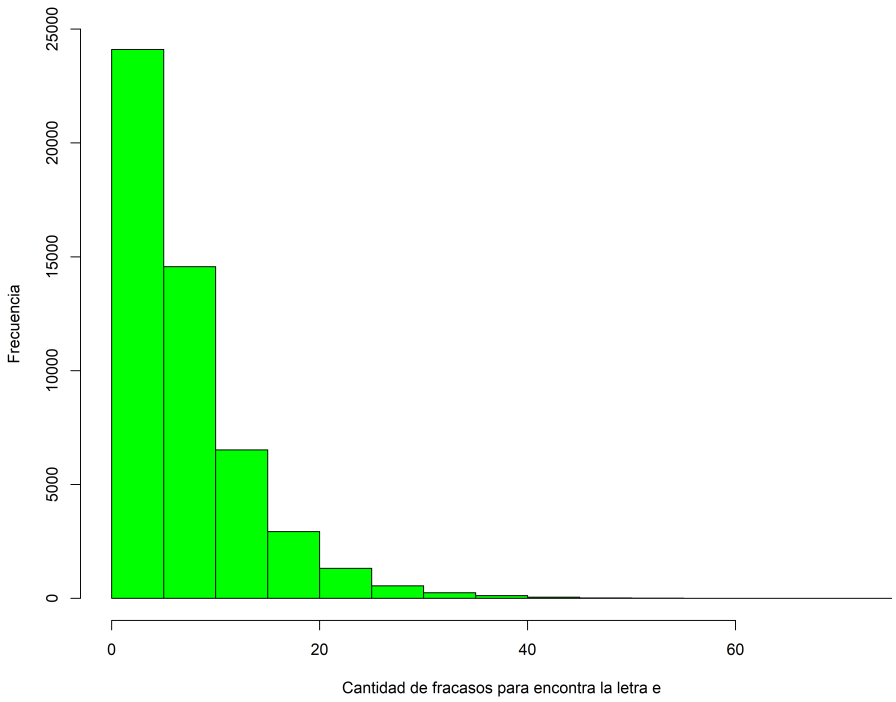

Figura 3: Distribución geométrica usando como éxito la letra con mayor frecuencia en el libro

ejemplo la letra y palabra con mayor frecuencia.

Para la distribución geométrica, se consideró el criterio de cuantos caracteres alfanuméricos hay antes de encontrar la letra con mayor frecuencia en todo el texto, en este caso, el éxito será encontrar la letra `e`. En la figura ref se puede observar como el criterio seleccionado sigue una distribución geométrica.

Para la distribución binomial se consideró que las repeticiones se formarían con los párrafos cuya longitud de palabras es igual a 9, ya que son los más frecuentes y como longitud de palabra igual a 2, ya que es la menor cantidad de caracteres que pueda contener este tipo de párrafos. Con esta simulación se pretende encontrar cuantas palabras con la longitud dada (éxitos) se encuentran en los párrafos con las características mencionadas anteriormente. En la figura 4 se puede observar como el criterio seleccionado sigue una distribución binomial. Cabe mencionar que como la muestra es muy pequeña, son muy pocos los datos obtenidos.

Finalmente, para la distribución binomial negativa, se consideró hallar la cantidad de intentos para completar 5 éxitos en todo el libro, se considera un éxito cuando la letra `h` tiene seguida la letra `e`. Se contempló este criterio ya que la palabra y artículo más frecuente es `hester` y `the`, respectivamente. En la figura 5 se puede observar como el criterio seleccionado sigue una distribución binomial negativa.

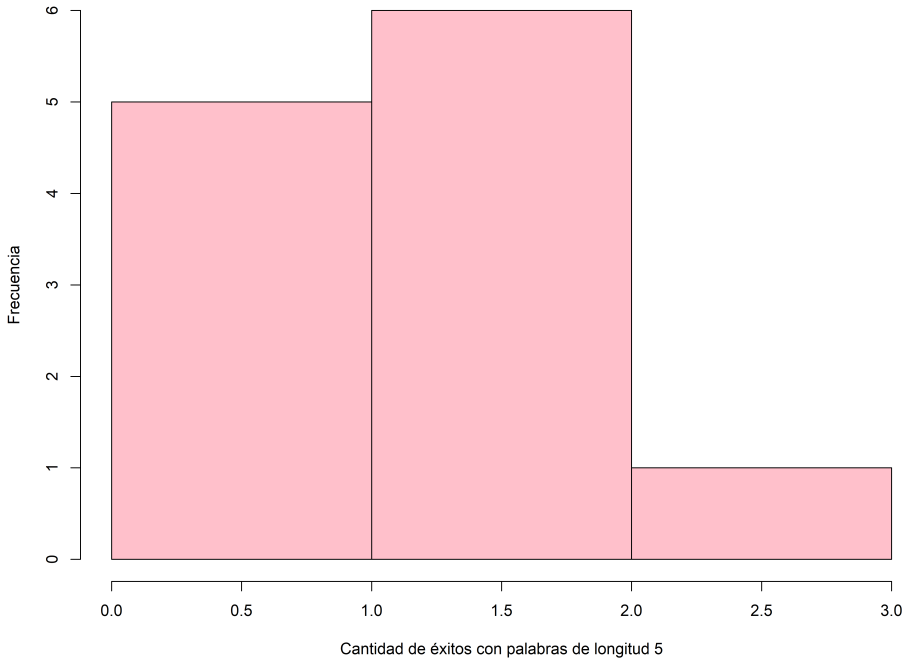

Figura 4: Distribución binomial usando como éxito la cantidad de palabras con una deterinada longitud

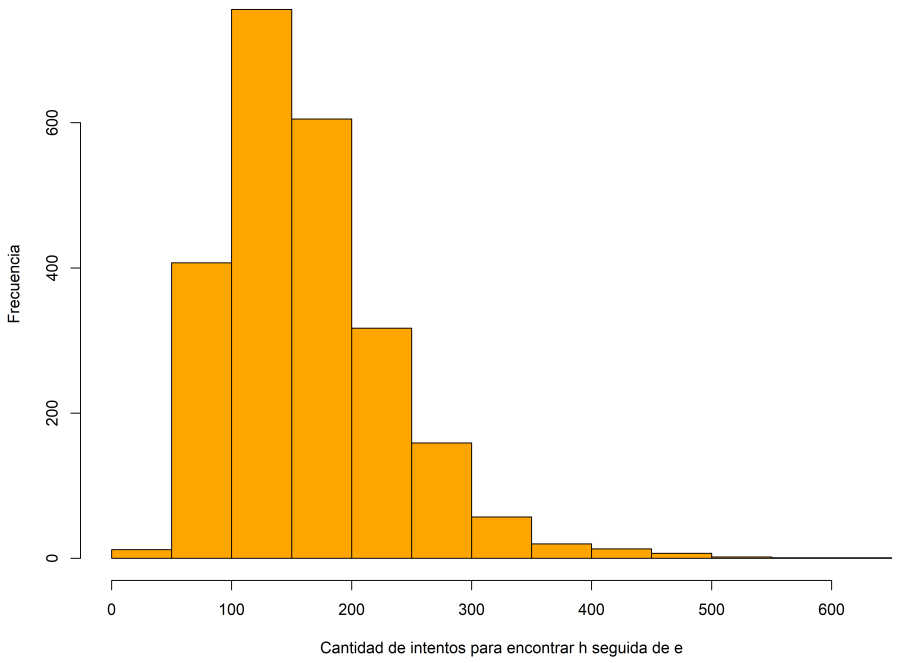

Figura 5: Distribución binomial negativa

Referencias

[1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea3, 2020.

[2] Hawthorne, Nathaniel. The Scarlet Letter. Recurso libre, disponible en http://www.gutenberg.org/ebooks/25344.

[3] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.